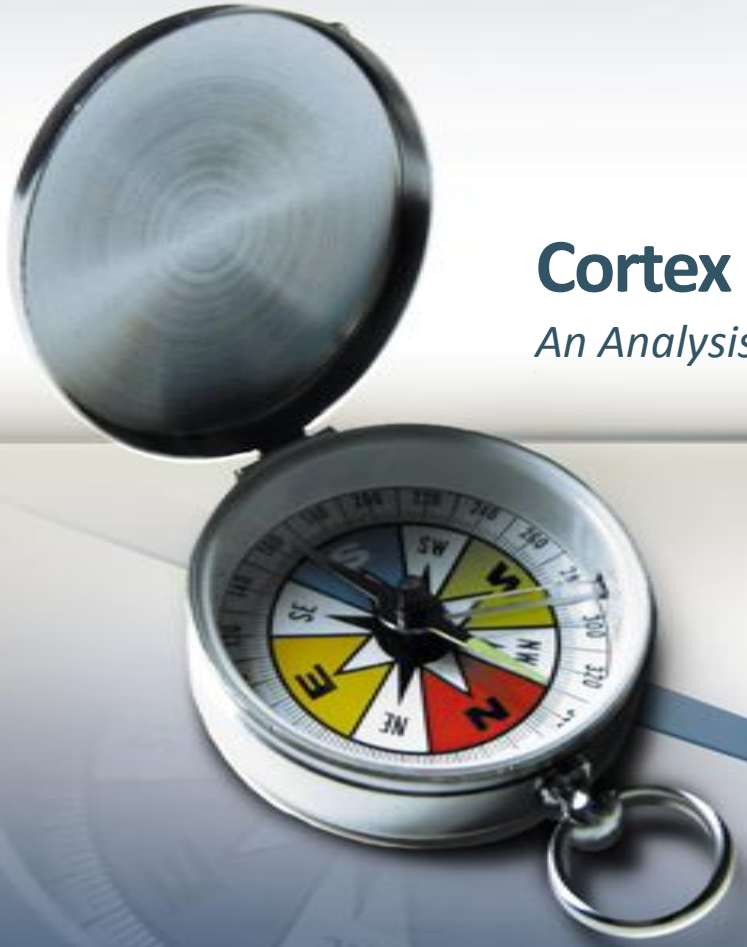
Cortex Intelligence at TAC 2010

An Analysis of The Cortex Method at TAC 2010 KBP Slot-Filling



cortex
intelligence

Agenda

- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

...but first, a little bit about Brazil, and Cortex

Growth

Currently the world's ninth **economy**, projected to become the **fifth** by **2016**

Cortex

Cortex Intelligence - reference in Portuguese text mining

Olympic City

Brazil was chosen to host the **2014 World Cup** and Rio de Janeiro was chosen to host the **2016 Olympic Games.** An endeavor bringing in an estimated **R$1.3 trillion** in new investments.

-source: SOBRATEMA

Social Media

Currently the **second largest** Twitter user base worldwide, nearing 10% of all users and 7% of all tweets. Over **400%** increase in Facebook userbase in 2010.

- source: sysomos.com

Microsoft BizSpark One

...but first, a little bit about Brazil, and Cortex

Key Facts About Cortex

- International reference in Text Mining and Web Semantics (1st place in text mining context for Portuguese)
- Today, this technology is used in market intelligence software and services for large corporations
- Based in Rio, 47 employees, over US$ 3 million in revenues and ~90% annual growth rate
- Key customers include some of the largest companies in Latin America in segments as banking, e-commerce, chemicals, mining, oil & gas, etc
- No funding from VC yet, only bootstrap and government grants for text mining research so far
- 1st company selected in Latin Am. by Microsoft for BizSpark One, its startup acceleration program

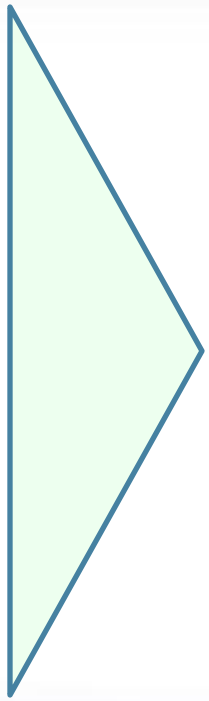
Agenda

- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

Introduction

Strategy

- ✓Evidence-based entity extraction
- ✓Linguistic rules
- ✓Flexible knowledge representation
- ✓Semantic constraints
- ✓Highly coupled interaction between complementary systems.



SIGNIFICANT RESULTS

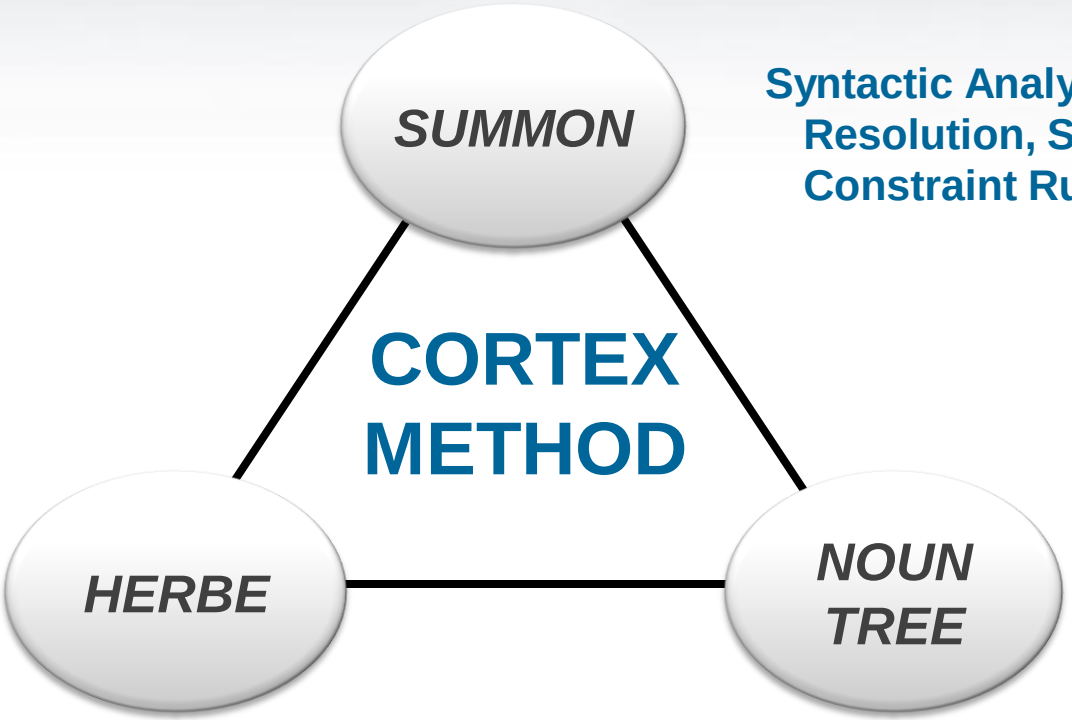
Top scoring system for regular slot-filling

Performance comparable to LDC's annotation on surprise slot-filling with limited human intervention

Agenda

- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

The Three Systems



Syntactic Analysis, Anaphora Resolution, Semantic Constraint Rules, Slot-Filling

Evidence and Resource-Based Entity Extraction

Flexible, Dynamic, Hierarchical Representation

HERBE – Heuristic Evidence and Resource-Based Extraction

cortex intelligence

HERBE is a Pipeline-Based Entity Extractor

COGNITIVE ARCHITECTURE – ***based on over 10 years of Dr. Christian Aranha's research and experience.***

IDENTIFICATION – ***The determination of the initial and final offsets of an entity or term***

CLASSIFICATION – ***The label ascribed according to an internal ontology***

HERBE's Responsibilities to Its Client Systems

- ❑**Identification**
- ❑**Collocation Pinpointing**
- ❑**Entity Chunking**
- ❑**First-level classification**
- ❑**Acronym pinpointing**
- ❑**First level Disambiguation (sentence boundaries, paragraph boundaries, punctuation roles)**
- ❑**Normalization**

Secondary Tasks

Initial co-reference resolution

Grammatical classification

Microsoft BizSpark One

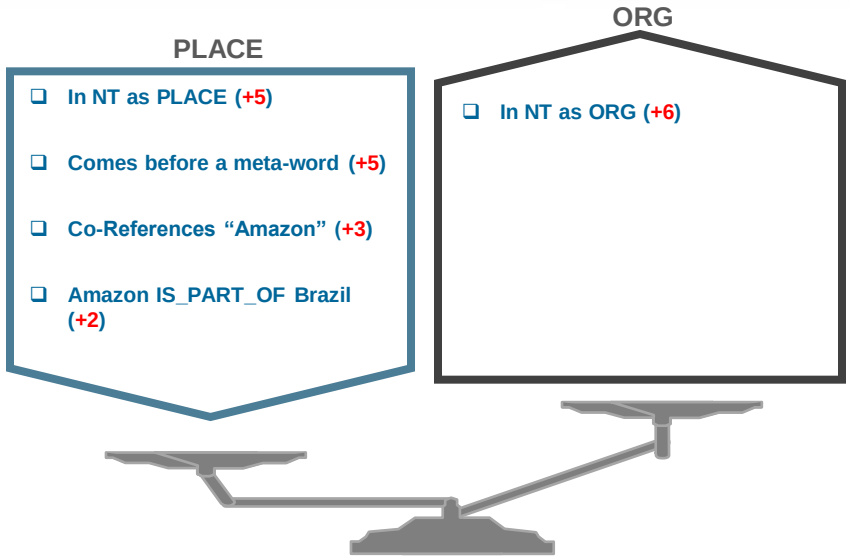
HERBE – Heuristic Evidence and Resource-Based Extraction

“Brazil will boost natural gas output from the *Amazon* region to reduce its dependence on Bolivian fuel.”

Evidence Accumulation

- ❑In NT as PLACE
- ❑Comes before a meta-word
- ❑Co-References “Amazon”
- ❑Amazon IS_PART_OF Brazil
- ❑In NT as ORG

Tally



Result

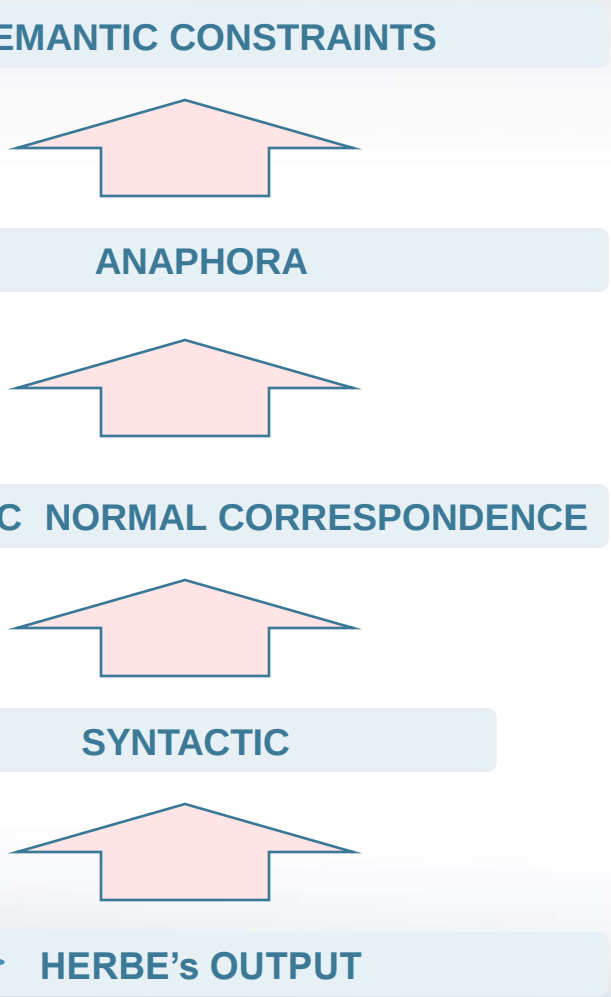
AMAZON = PLACE

SUMMON – Summoning Relations

SUMMON's Responsibilities

- ❑ **Syntactic Analysis**
- ❑ **Syntactic Correspondence Normalization**
- ❑ **Extended Co-reference and Anaphora Resolution**
- ❑ **Application of Semantic Constraint Rules**
- ❑ **Event Extraction**
- ❑ **Slot-Filling**

SUMMON consumes the word sequence generated by HERBE (or previous versions of Cortex's EE systems), and transforms it into a word graph, in which every word is connected to others over which it exerts a syntactic relation. This system is not phrase-based, but herein connected words possess a syntactic relation to one another.



Layered approach

Incremental normalization and abstraction

Advanced co-reference resolution

SUMMON – Summoning Relations

SUMMON's Anaphora Resolution

Pronominal Resolution

"***They're certainly following the predicted pattern, building Alice Dellal up more and more with stories like this …****(and further down the page)* ***…One of her aunts, Suzy, died from a heroin overdose while studying in Paris.***"

Name-part Resolution

"***New York City Opera has commissioned American composer Charles Wuorinen*** *(and further down the page)* ***Wuorinen, 70, said in a statement.***"

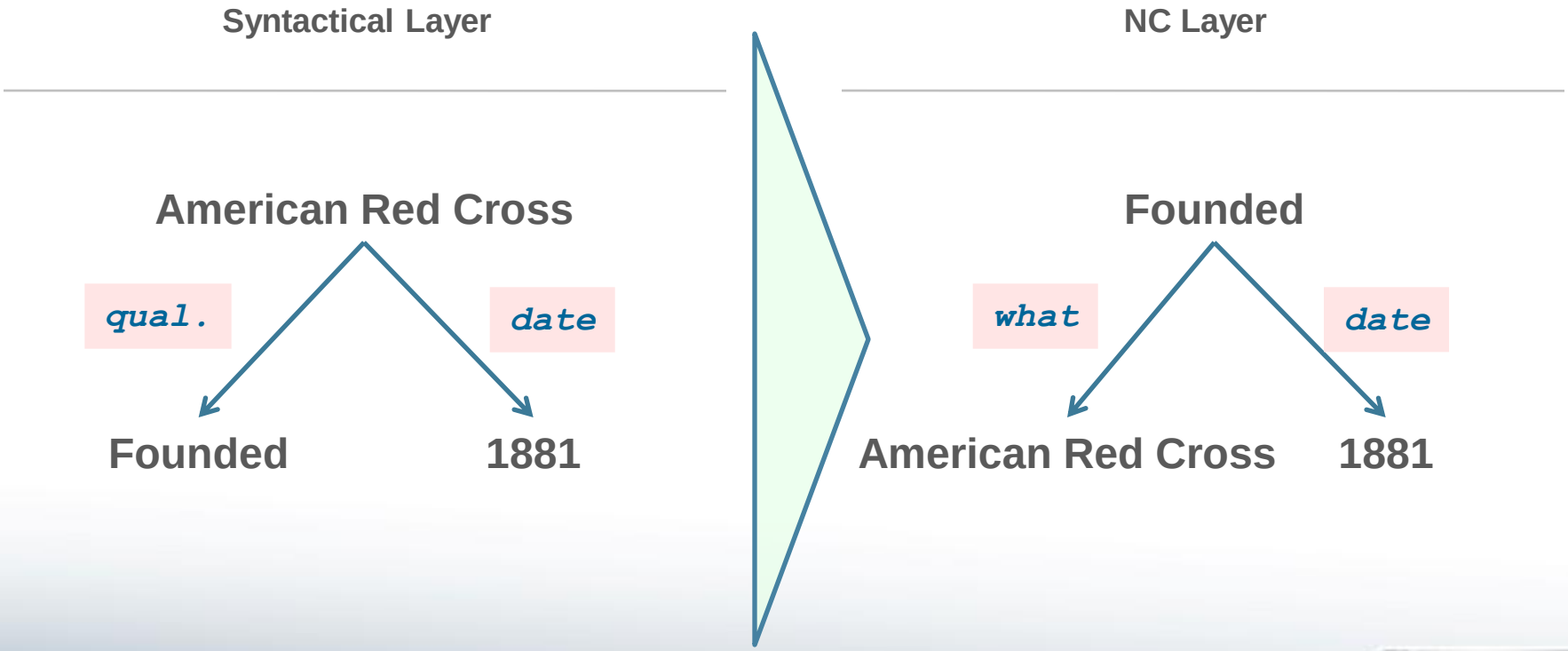
Semantic Constraint Resolution

"***Madonna won a court battle Monday against a British tabloid that published pictures recently of her wedding eight years ago. The singer's adopted 3-year-old son, David Banda ...***"

SUMMON – Flexible Representation

Syntactic Normal Correspondence

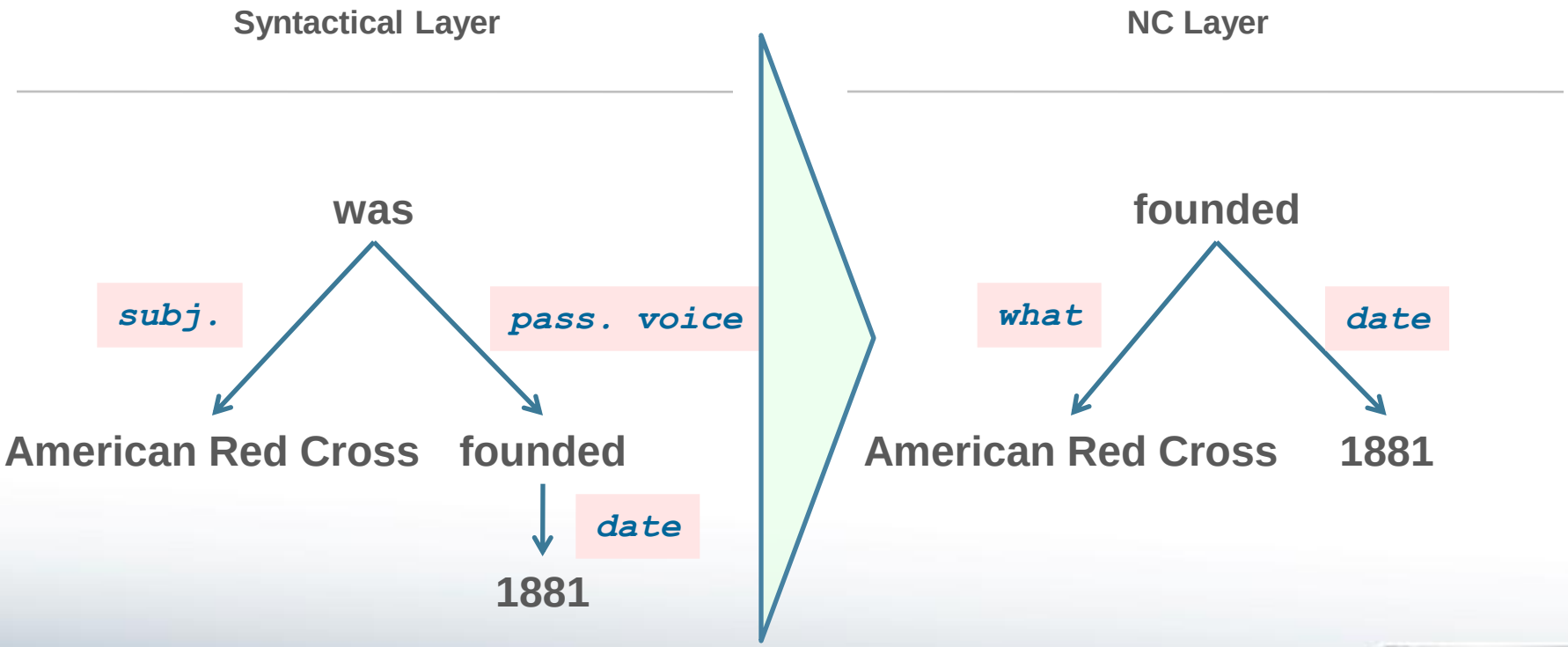
"Founded in 1881, the American Red Cross ..."



SUMMON – Flexible Representation

Syntactic Normal Correspondence

"The American Red Cross was founded in 1881 …"

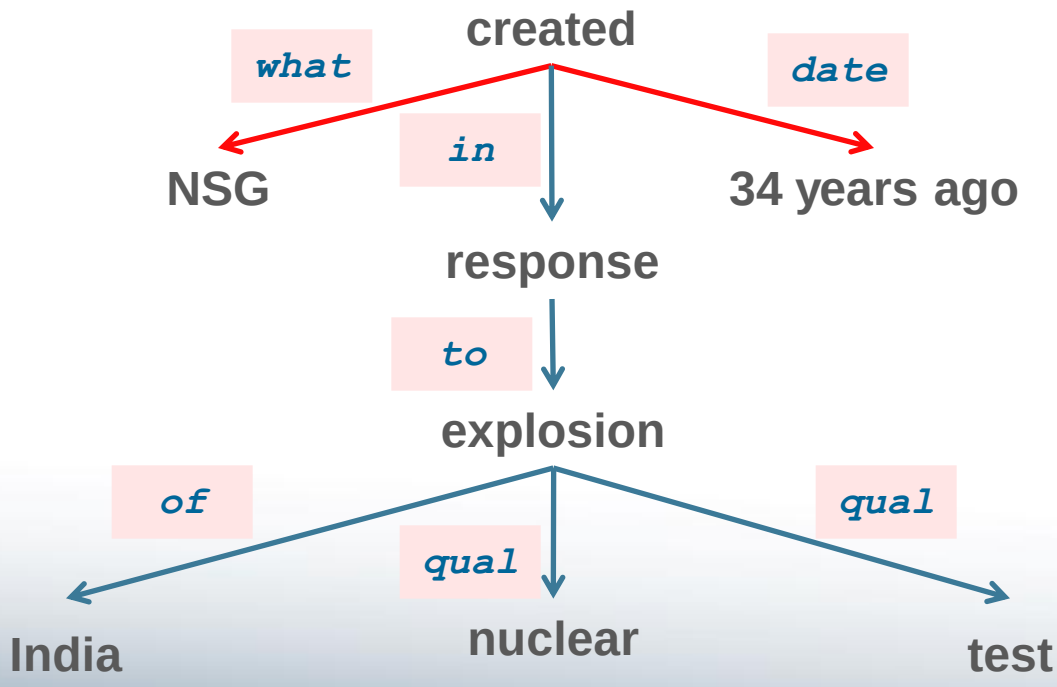


SUMMON – Flexible Representation

Syntactic Normal Correspondence

"The NSG was created in response to India's nuclear test explosion 34 years ago"

NC Layer



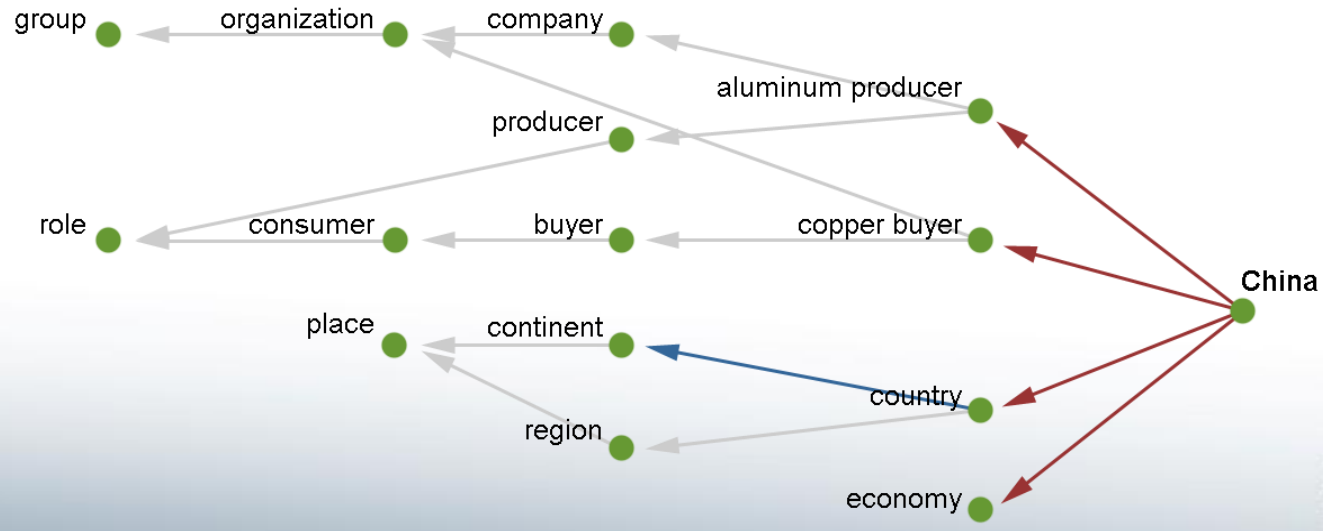
NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- ❑Directed
- ❑Acyclic
- ❑Multiple Parents

- ❑Main Lexicon and Knowledge Repository for HERBE and SUMMON
- ❑Has Internal Logic to Maintain Its Own Consistency

Example of NOUN TREE's Relations

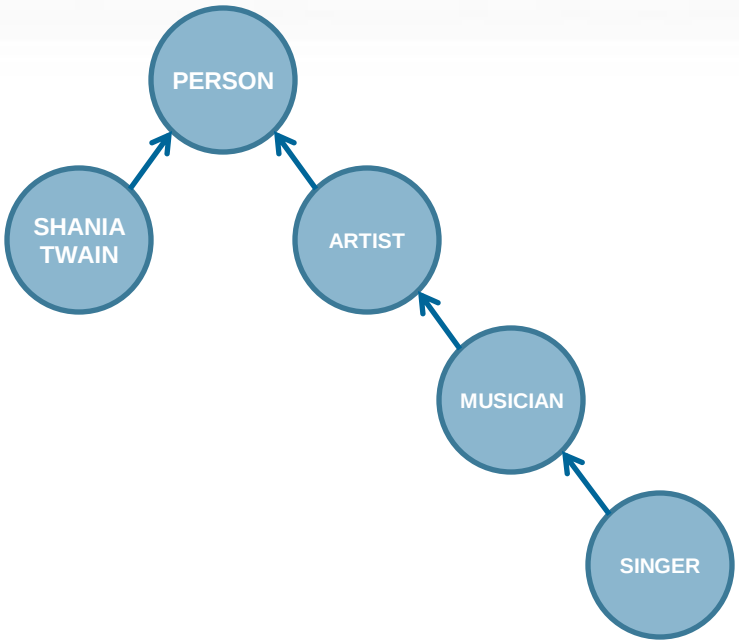


NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- "-Is Shania Twain a person?"
  "-Yes."
- "-Is Shania Twain an artist?"
  "-No."
- "-Is Shania Twain a musician?"
  "-No."
- "-Is Shania Twain a singer?"
  "-No."

Example of NOUN TREE's Knowledge Accumulation

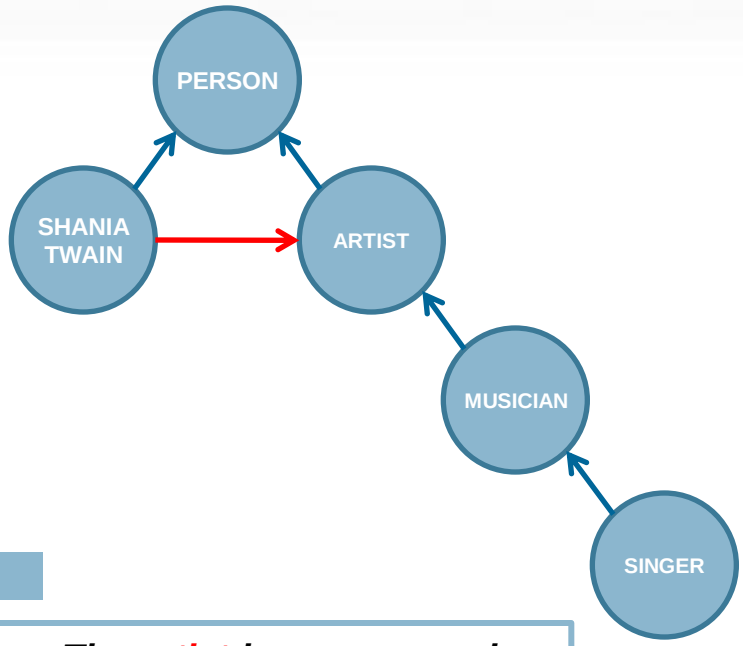


NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- "-Is Shania Twain a person?"
  "-Yes."
- "-Is Shania Twain an artist?"
  "-No."
- "-Is Shania Twain a musician?"
  "-No."
- "-Is Shania Twain a singer?"
  "-No."

Example of NOUN TREE's Knowledge Accumulation



Passage Example

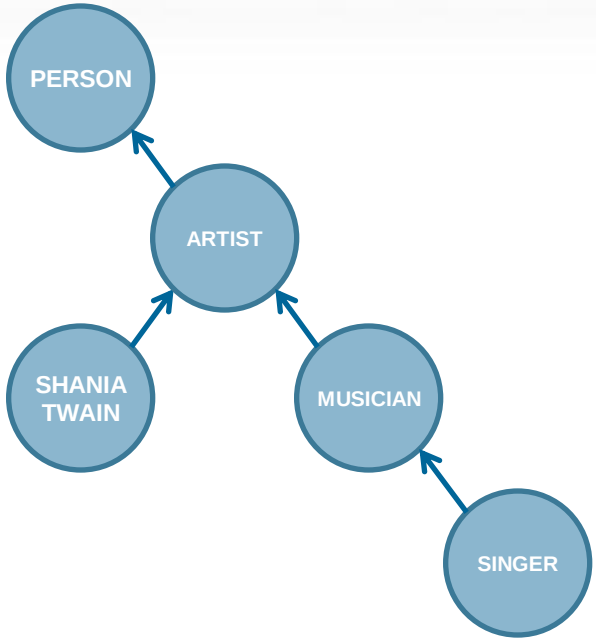
"Shania Twain has released a new album. The artist has expressed profound happiness with this latest work"

NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- "-Is Shania Twain a person?"
  "-Yes."
- "-Is Shania Twain an artist?"
  "-Yes."
- "-Is Shania Twain a musician?"
  "-No."
- "-Is Shania Twain a singer?"
  "-No."

Example of NOUN TREE's Knowledge Accumulation

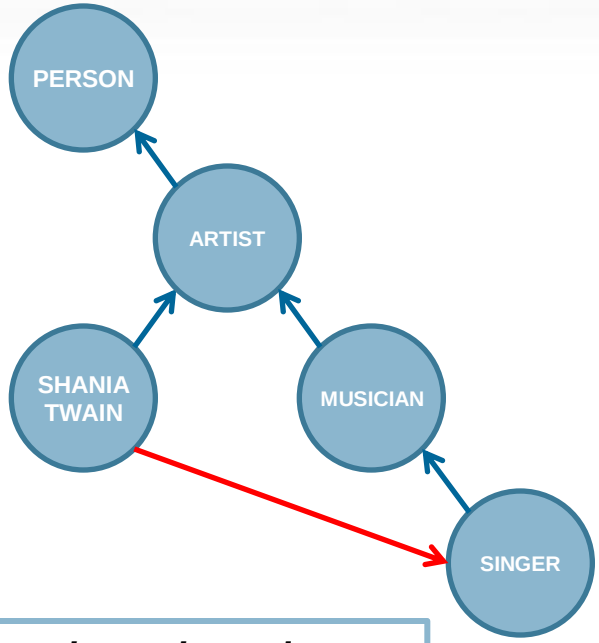


NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- "-Is Shania Twain a person?"
  "-Yes."
- "-Is Shania Twain an artist?"
  "-Yes."
- "-Is Shania Twain a musician?"
  "-No."
- "-Is Shania Twain a singer?"
  "-No."

Example of NOUN TREE's Knowledge Accumulation



Passage Example

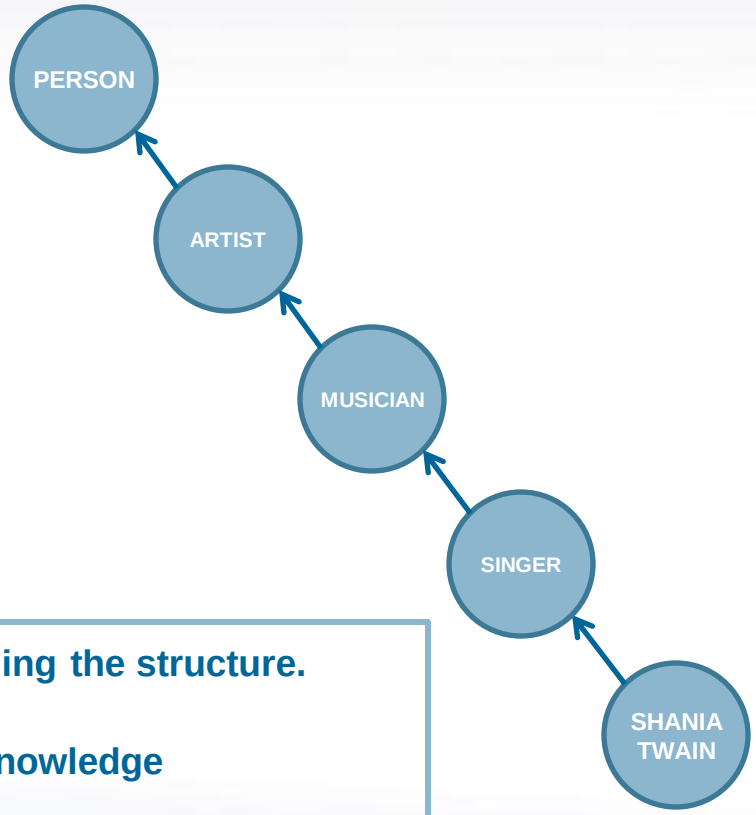
"Shania Twain, the popular American country singer, has released a new album."

NOUN TREE – Flexible Representation

NOUN TREE's characteristics

- "-Is Shania Twain a person?"
  "-Yes."
- "-Is Shania Twain an artist?"
  "-Yes."
- "-Is Shania Twain a musician?"
  "-Yes."
- "-Is Shania Twain a singer?"
  "-Yes."

Example of NOUN TREE's Knowledge Accumulation



Advantages

- Knowledge is accumulated without over-burdening the structure.
- Dynamic structure changes allow flexibility in knowledge representation
- API for Interaction with SUMMON and HERBE to answer specific 'questions' correctly without loss of semantic plausibility.

Agenda

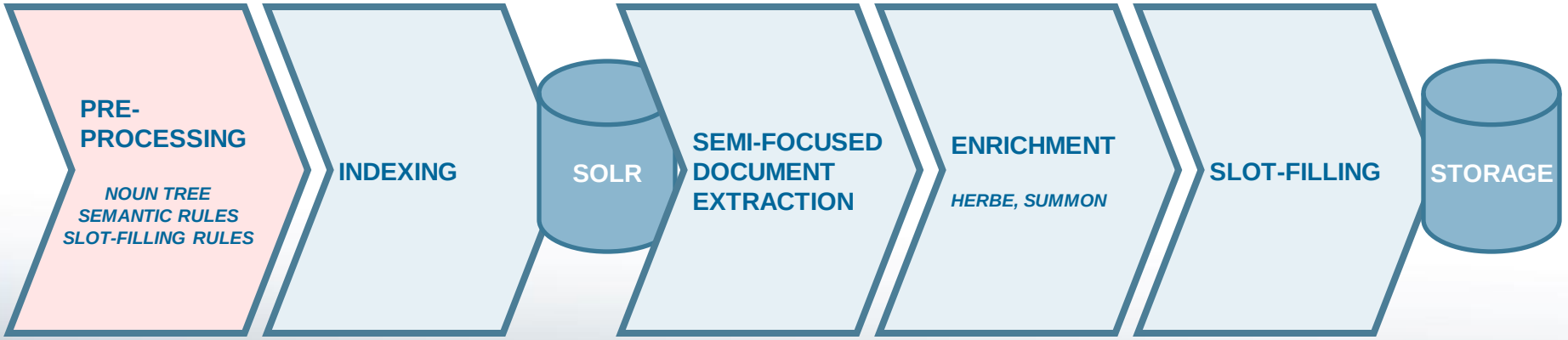
- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

The Application - Preparing

Pre-processing Step of the Slot-Filling App

- Used the Cortex English Corpus (over 3 million documents), and the ***news and blog*** subset of TAC Corpus for:
  - **Noun Tree enrichment** - new relations and new entities
  - **Constraint building** - finding the best semantic constraints for slot-filling rules in each slot type

- **Transliterations**: Dzokhar Dudayev, Djokhar Dudayev, Dzokar Dudaev, …?
- The variant transliteration system was a hybrid of a **phonetic root substitution** system and a modified fast **Levenshtein** distance calculator.
- Unhandled variances in transliteration affected both **recognition** and **classification** in later steps.



Agenda

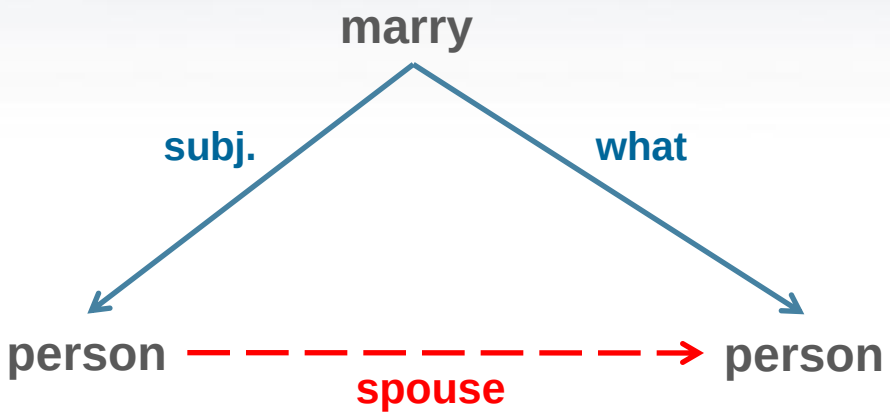
- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

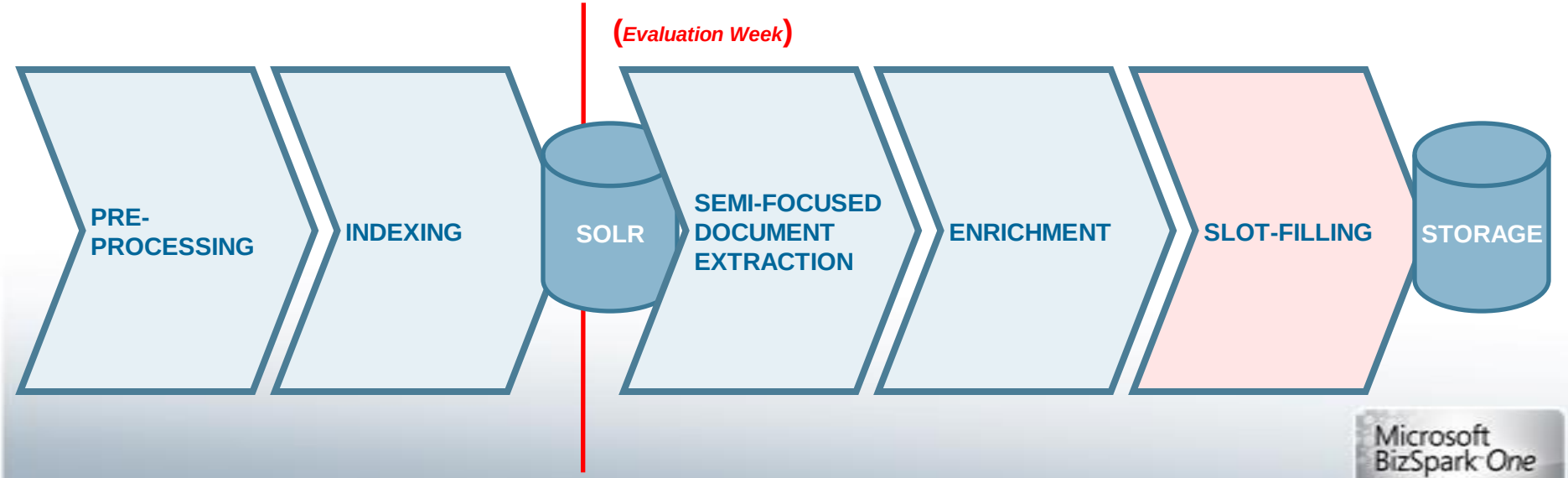
Regular Slot-Filling

Highlights of The Slot-Filling Application

- New layer over the **SUMMON** system
- Slots were populated according to semantic **rule-patterns** which had to match some **sub-graph** of the text
- Just over **200 slot rules**

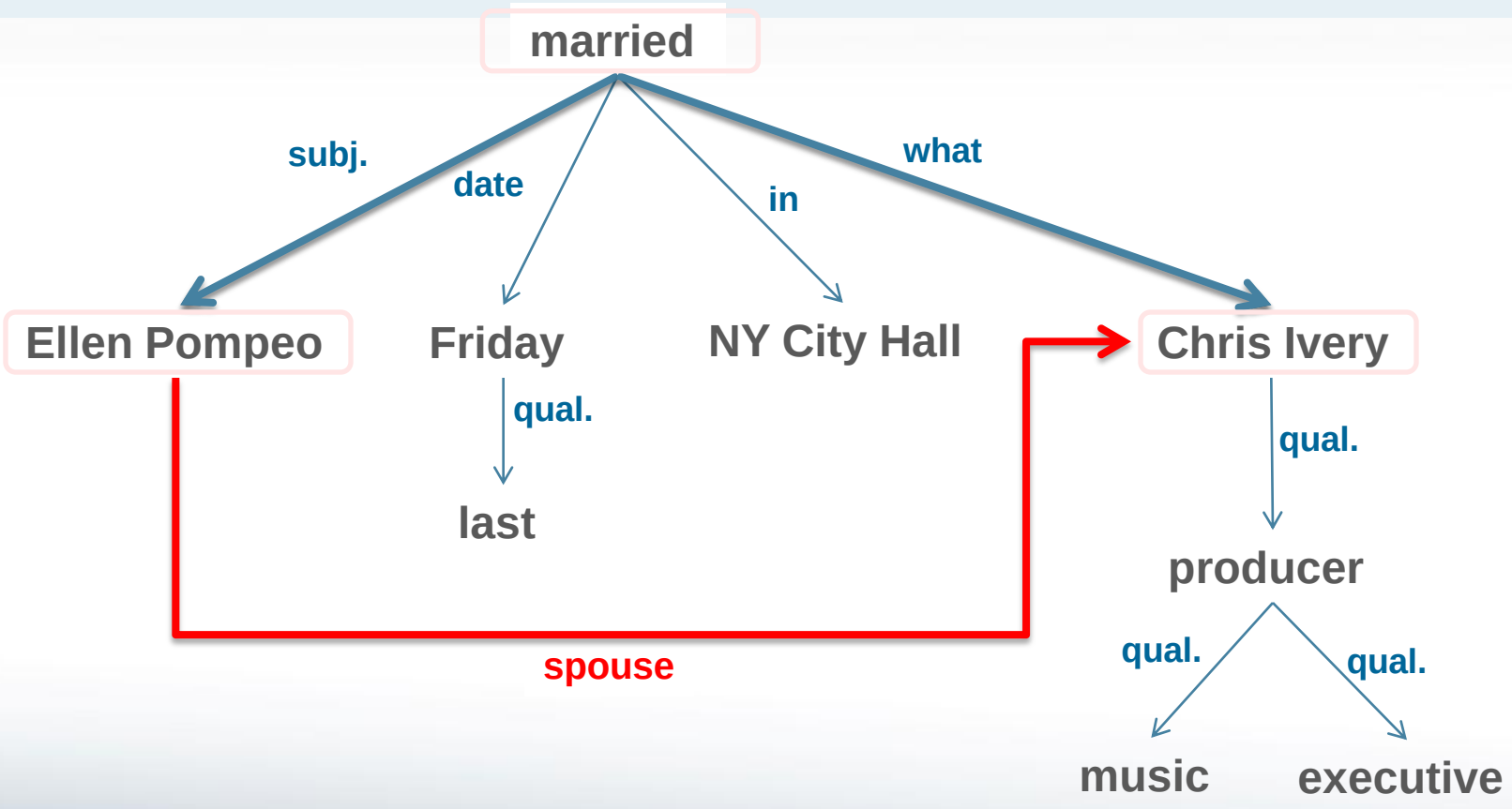
Example Rule





Regular Slot-Filling

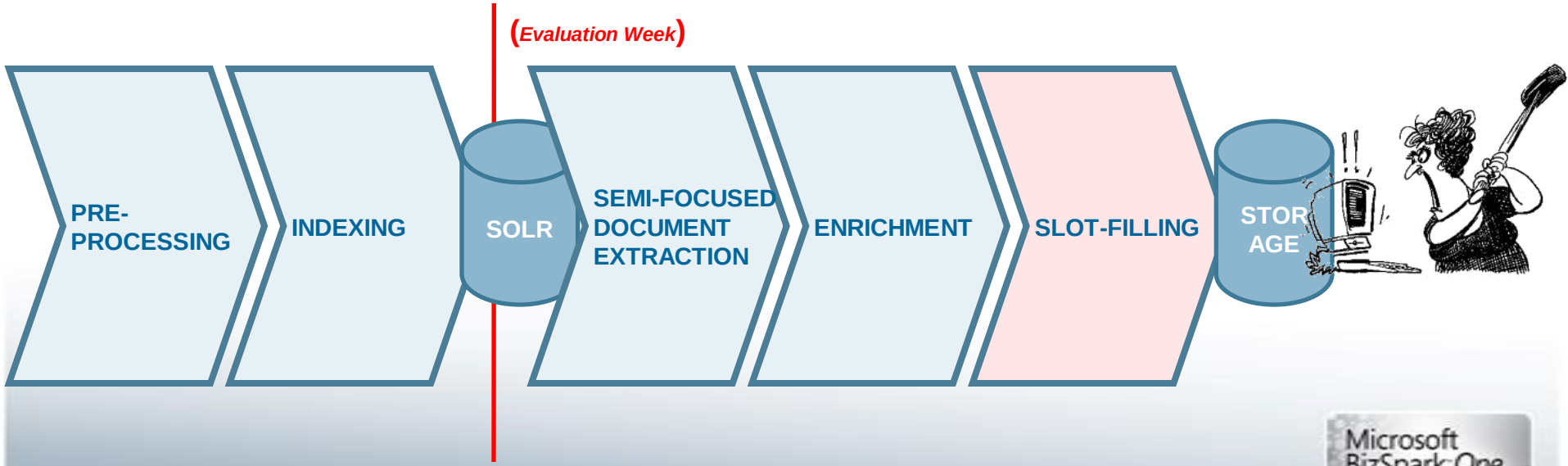
"*Ellen Pompeo got married last Friday in NY City Hall with the music executive producer Chris Ivery*"



Surprise Slot-Filling

Highlights of The Slot-Filling Application

- **Absence** of slot-specific semantic constraint rules
- Higher **recall** – *much* lower **precision**
- Direct human Intervention – **pass-fail** interface
- In essence: a highly effective hybrid **suggestion system**



Agenda

- Introduction
- The Three Systems
- The Application
- Slot-Filling
- Analysis and Conclusion

Analysis and Conclusions – Regular Slot-Filling

Regular Slot-Filling	LDC	Top Score	2nd Score	Median	Cortex
PRECISION	0.7013802	0.667996	0.6655173	0.21414538	0.667996
RECALL	0.54061896	0.64796907	0.18665378	0.10541586	0.64796907
F-MEASURE	0.6105953	0.6578301	0.29154077	0.14128321	0.6578301

CONCLUSIONS

- **Strong evidence to the robustness of the combination of syntactic analysis with semantic constraints in textual enrichment.**
- **Well-tuned semantic constraints are key**

Analysis and Conclusions – Surprise Slot-Filling

Regular Slot-Filling	LDC	Top Score	2nd Score	Median	Cortex
PRECISION	0.8531746	0.69215685	0.52360517	0.5032258	0.69215685
RECALL	0.42574257	0.6990099	0.24158415	0.15445544	0.6990099
F-MEASURE	0.5680317	0.69556653	0.3306233	0.23636363	0.69556653
TIME	N/A	99	34	11	99

CONCLUSIONS

- **Strength that syntactic-semantic constraint rules and flexible representation provide to an automated system**
- **Again …well-tuned semantic constraints are *key*!**

THANKS.

Cortex Intelligence

www.cortex-intelligence.com
daniel.chada@cortex-intelligence.com
Tel: (55 21) 3282-3150

Rua da Assembléia, 10/3711
Centro, Rio de Janeiro, RJ